Automated Classification for Visual-Only Post-Mortem Inspection of Porcine Pathology

McKenna, Stephen; Amaral, Telmo ; Kyriazakis, Ilias

Published in:
IEEE Transactions on Automation Science and Engineering

DOI:
10.1109/TASE.2019.2960106

Publication date:
2020

Document Version
Peer reviewed version

Automated Classification for Visual-Only Post-Mortem Inspection of Porcine Pathology

Stephen McKenna, *Senior Member, IEEE*, Telmo Amaral, and Ilias Kyriazakis

***Abstract*—Several advantages would arise from the automated detection of pathologies of pig carcasses, including avoidance of the inherent risks of subjectivity and variability between human observers. Here, we develop a novel automated classification of two porcine offal pathologies at abattoir: a focal, localized pathology of the liver and a diffuse pathology of the heart, as cases in point. We develop a pattern recognition system based on machine learning to identify those organs that exhibit signs of the pathology of interest. Specifically, deep neural networks are trained to produce probability heat maps highlighting regions on the surface of an organ potentially affected by a given condition. A final classification stage then decides whether a given organ is affected by the condition in question based on statistics computed from the heat map. We compare outcomes of automated classification with classification by expert pathologists. Results show classification of liver and heart pathologies in agreement with an expert at levels comparable to, or exceeding, inter-expert agreement. A system using methods such as those presented here has potential to overcome the limitations of human-based abattoir inspection, especially if this is based on visual-only inspection, and ultimately to provide a new gold standard for pathology.**

***Index Terms*—Agriculture; Food industry; Machine vision; Neural network applications**

NOTE TO PRACTITIONERS

The motivation for this paper reflects the current requirement for visual only inspection of livestock carcasses at slaughter houses and the need to provide a gold standard for recognition of carcass pathologies. Visual only inspection is motivated by the need to reduce cross contamination between carcasses by manual palpation, but this leads to substantial variability in detection accuracy both within and between inspectors. This has significant public health implications. Here we present a system that comprises hardware to capture images of pig offal and software to analyse those images and identify cases of liver milk spots and hearts affected by pericarditis. It can classify high proportions of offal with accuracy comparable to that of veterinarians with extensive experience in pig pathology, thus demonstrating the potential to overcome the limitations of human-based abattoir inspection (especially if it is visual-only) and ultimately to provide a new gold standard. Our work is the first to address the automation of pig offal inspection, thus shedding light on the challenges associated with both appropriate image capture and successful image analysis, such as the need to cope with wide variations in the appearance of both normal and diseased organs, as well as different types of lesions and their impact on how much effort is required from experts in order to produce data needed to train the system. Future directions of work should include extending the system to identify more pathologies and implementing a real-time system to cope with production line speed.

T. Amaral was with Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.

I. Kyriazakis is with Agriculture, School with Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.

S. McKenna is with CVIP, Computing, School of Science and Engineering, University of Dundee, Dundee DD1 4HN, UK. e-mail: (see http://staff.computing.dundee.ac.uk/stephen/).

Manuscript submitted November, 2017. Resubmitted September 2018.

I. INTRODUCTION

THE main purpose of meat inspection is to assure consumers about the safety, hygiene and nutritional value of their food. Meat inspection can also help detect and prevent public health hazards such as food-borne pathogens or chemical contaminants in food of animal origin [1]. Under more detailed examination [2] carcass inspection post-mortem can provide information about underlying conditions that may affect pig performance on farm, but do not manifest as clinical disease, as in the case of Holt et al. [3] who found associations between prevalence of respiratory lesions and respiratory pathogens in the herd. Such information can assist the development of on farm health control plans.

Because meat inspection relies heavily on human observation, it carries inherently the risks of subjectivity and variability between observers who have limited time to observe individual carcasses within an abattoir production line. In the case of pig meat inspection, the situation is further complicated by two additional factors: (i) there is a move towards visual only inspection of pig carcasses and offal, without any involvement of palpation, on the grounds of minimizing the risk of cross contamination between carcasses. This has been the case in the European Union since 2014 [1]. (ii) There is disagreement in the evaluation schemes of pathologies between countries and states, as has been shown by Steinmann et al. [4] for the detection of respiratory lesions of pig carcasses within the EU. The latter authors have found that current evaluation and recording of lesions by authorized meat inspectors are not reliable and produce signicant inter-rater disagreement. These limitations may be overcome through the use of automated inspection systems, with the additional advantage of detecting a greater number of pathologies than currently observed by meat inspectors, due to time limitation [3].

In this paper we develop a methodology for the visual-only, automated classification of pathologies of pig carcasses at abattoir. We concentrate on the challenge of detecting pathologies on pig red offal [5], as most conditions that pose

health hazards are associated with that part of the pig [3]. Pig red offal is made up of inter-connected organs that belong to the pig's non-digestive tract, primarily the heart, lungs, diaphragm and liver. This group of organs is termed a *pluck*. In the abattoir, the pluck from each pig hangs from a hook in the production line. Fig. 1 shows an example pluck image and the areas occupied in it by each organ. We focus on the detection of pathologies on pig livers and pig hearts. The main liver pathologies involve the presence of lesions that result from the infiltration of lymph cells in reaction to parasite infection, and are commonly encompassed by the term *milk spots*[1]. Pericarditis is an inflammation of the pericardium, the membrane that encloses the heart, which becomes thickened by fibrous tissue and adheres to the heart [6]. From a technical perspective, these two conditions are of interest as they represent the two extremes of what an automated detector may have to deal with: very localized lesions (milk spots) and very diffuse lesions covering most of the surface of the affected organ (pericarditis).

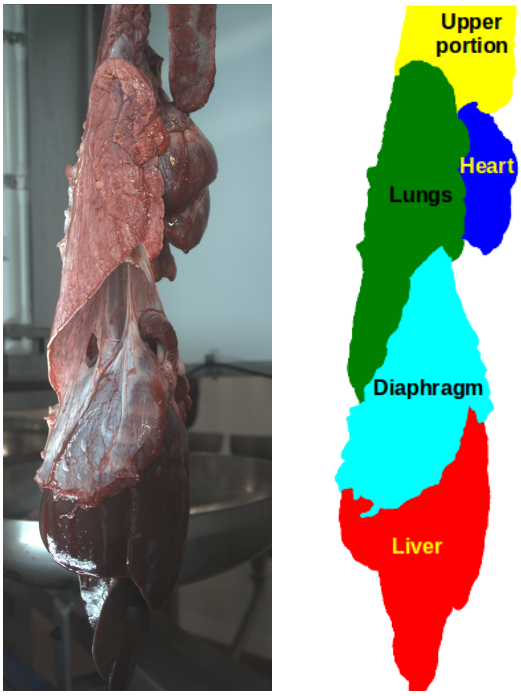
The method we propose uses deep convolutional neural networks [7] to assign pathology probabilities to image locations. The resulting image of probabilities can be visualised as a heat map that highlights regions on the surface of an organ potentially affected by a given condition. A final classification stage then decides whether a given organ is affected by the condition in question, based on statistics computed from that organ's heat map. Results show that this approach enables classification of livers as to the presence of milk spots, and hearts as to the presence of pericarditis, in agreement with an expert at levels comparable to, or exceeding, inter-expert agreement. Our work is the first to address the automation of pig offal inspection, thus providing insight into the technical challenges associated with both capture and analysis of these types of image.

II. Related Work

Automated image analysis has been quite widely applied to the assessment of meat quality and safety. Ma et al. [8] provide a short review of some of these applications. Craigie et al. [9] review the use of image analysis for evaluation of beef carcasses specifically. Xiong et al. [10] review the range of non-invasive imaging technologies that have been applied. These include thermal, hyperspectral, fluorescene, magnetic resonance, x-ray, and ultrasound imaging.

The use of computer vision at abattoir has been investigated for inspection of limited pathologies of broiler chickens; specifically, a system for inspection of footpad dermatitis has been evaluated [11] and a method for classifying broiler livers has been proposed [12], [13]. Segmentation of organs was attempted using colour thresholding and active contours in the case of poultry viscera [14]. However, the use of modern machine learning methods is promising for such tasks and has been reported for poultry viscera [15] and in our own research on porcine offal [16]. Less recent work on poultry carcasses includes the detection of abnormal livers and hearts [17], diseased air sacs [18], and splenomegaly [19].

[1]http://www.nadis.org.uk/bulletins/ascariasis.aspx



(a) (b)

Fig. 1: (a) Example image of "pluck". (b) Regions occupied by different organs.

Computerised image analysis of pathology is extensively addressed in the field of human medicine where parallels might be drawn with the expansion of whole-slide digital pathology driving development of systems for high-throughput, systematic, and reproducible analysis of these large datasets [20], [21]. Analysis of macroscopic visible light images includes extensive work on polyp detection from colonoscopy [22], [23] and classification of skin lesions [24], [25].

III. Image Acquisition

Images used in this paper were captured by a solution that was designed by Hellenic Systems Ltd. It incorporated two area-scan colour cameras acquiring images from different viewpoints at high-resolution (4000× 5120 pixels) as exemplified in Fig. 2. Cameras and LED tube lighting were housed in a stainless steel canopy-shaped structure installed on the production line. Both cameras and tube lights were fitted with polarising filters which reduced the effect of specular reflections on the surface of the organs. The cameras were protected from the harsh environment using custom-made enclosures to IP68 standard with polycarbonate viewports. Cameras were triggered automatically using a PLC controller and proximity switches each time a pluck, hanging from a hook, passed through the canopy structure. Images were then processed by local control systems for sortation and storage before being passed to a central server.

IV. Pathology Identification by Experts

Two veterinarians with expertise in pig pathology and previous experience of working on the British Pig Health

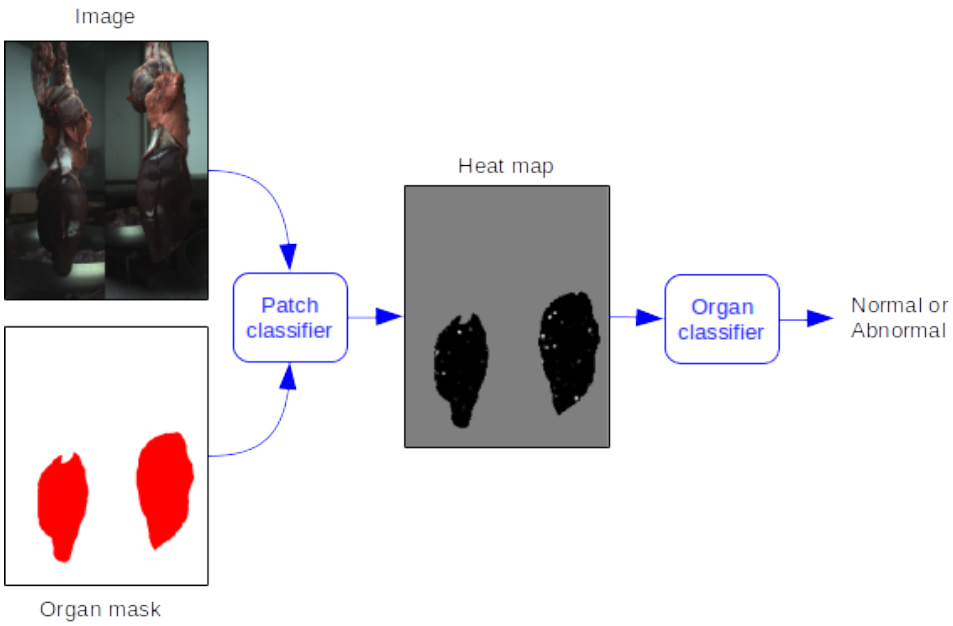


Fig. 2: Overview of pattern recognition system for organ classification

Scheme (KY and JW) identified and labelled those images in which pathologies of interest were visible. KY had 14 years of experience in pig health, including pathology and production; JW had 39 years of experience in veterinary practice, specializing exclusively in pigs, including routine slaughter surveys and involvement in BPHS surveys from 2005 until 2016.

Both images (two viewpoints) per pluck were inspected. Each pluck's images were inspected for a range of conditions. Images were labelled by categorising the organs into three classes: positive (showing signs of the condition in question), negative, or invalid (either due to unfocused or badly lit image, or because it was impossible to determine whether or not the condition was present). Here, we focus on pericarditis and milk spots. Images with no conditions (normal cases) were sampled from the available data to form reasonably balanced datasets.

The labelling process was performed by viewing images one at a time on ultra-high resolution monitors (Dell P2715Q) with in-plane switching (IPS) technology which allowed the monitors to be positioned in portrait orientation while preserving a good horizontal viewing angle. This enabled a full resolution image to be viewed without the need for zooming in and out to observe details.

The time taken to inspect images varied with the number and combination of lesions present on a pluck. It averaged approximately one image per minute. Both veterinarians took short breaks from the task after each session of approximately 45 minutes.

TABLE I: Confusion matrices for intra-observer and inter-observer reliability of two observers (coded KY and JW).

		KY_2		
		Neg	Pos	Inv
KY_1	Neg	190	3	2
	Pos	46	142	9
	Inv	0	0	0

(a) Liver (milk spots): intra-observer

		KY_2		
		Neg	Pos	Inv
KY_1	Neg	290	6	2
	Pos	8	188	4
	Inv	0	2	1

(b) Heart (pericarditis): intra-observer

		JW		
		Neg	Pos	Inv
KY_1	Neg	46	2	2
	Pos	25	88	1
	Inv	0	0	0

(c) Liver (milk spots): inter-observer

		JW		
		Neg	Pos	Inv
KY_1	Neg	115	0	1
	Pos	47	60	6
	Inv	1	0	0

(d) Heart (pericarditis): inter-observer

A. *Reliability of Identification by Experts*

The reliability with which the expert veterinarians classified the images was investigated. Repeatability (intra-observer variation) was analysed by comparing classifications made by KY on two separate occasions. Specifically, KY classified 392 livers and 501 hearts for a second time after a minimum of two months had elapsed, with order of presentation re-randomised and KY blinded to her initial classifications. Tables I(a) and I(b) give the intra-observer confusion matrices for liver classification (milk spots) and heart classification (pericarditis), respectively. It can be observed from these

matrices that heart classification had a high intra-observer repeatability while that of liver classification was somewhat lower. In the latter case, the observer appears to have shifted propensity to assign negative versus positive labels.

Reproducibility in terms of inter-observer variation was analysed by comparing classifications made by KY and JW. Specifically, JW classified 164 livers and 230 hearts from those data that had been classified twice by KY. JW was blinded to KYs classifications. Tables I(c) and I(d) give inter-observer confusion matrices for liver classification (milk spots) and heart classification (pericarditis), respectively. Both inter-observer classification matrices suggest that JW uses the negative classes more frequently than KY. Agreement on how to use the invalid class seems weak, although such images were relatively rare in this dataset.

These qualitative remarks are supported by descriptive statistics and statistical test results computed from the confusion matrices and provided in Table II. The specific agreement, i.e., the proportion of agreement specific to each class, is given in the final column. We tested for marginal homogeneity, i.e. lack of significant differences between row proportions and column proportions in the confusion matrices. Marginal homogeneity indicates that classes are being used with similar propensity. The Bhapkar test (df=2) was used to test for marginal homogeneity for all three classes simultaneously. Differences were significant in all experiments except for intra-observer variation on heart classification. Marginal homogeneity was also tested for each class separately using McNemar tests. (An exact test was used in some cases where the number of images in disagreement was small). Again significance was obtained (after Bonferroni correction) in all experiments except intra-observer variation on heart classification.

V. Automatic Classification

We developed a pattern recognition system based on machine learning to automatically classify organs according to whether or not they exhibit signs of a pathology of interest. Different pathologies vary greatly in their appearance, spatial distribution and extent. Fig. 2 shows an overview of the system illustrated using the example of milk spots. It has two main stages. The first stage classifies small patches of image, computing for each of them the probability that that patch is centred within a region of pathology. This is achieved using a deep convolutional neural network (CNN) trained on many patches. These probabilities can be visualized together as a heat map, highlighting potential regions of pathology. The second stage processes this map to compute an overall classification for the organ as either normal or pathological.

The architecture of the patch classifiers is shown in Fig. 3. This is a modification of the AlexNet architecture [26]. The patch classifier takes as input a high-resolution colour image patch of size 192×192. This is larger than nearly all liver milk spots. (A random sample of 50 milk spots were manually measured and had an average width of 70 pixels with standard deviation of 40 pixels). Image patches were extracted from the organ's image at locations on a grid spaced 24 pixels apart, both horizontally and vertically. Organ masks were

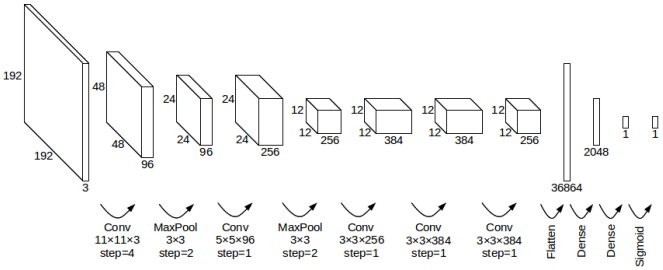


Fig. 3: Diagram of convolutional architecture used at the patch classification stage.

morphologically eroded to avoid extracting patches containing large portions of different organs or background. The CNN's first convolutional layer filters the patch using 96 kernels of size $11\times$ $11\times$ 3 and a step of 4 pixels, resulting in 96 planes of $48\times$ 48 values. These planes were subjected to a max-pooling operation using a $3\times$ 3 neighbourhood and a step of 2 pixels, resulting in 96 planes of $24\times$ 24 values. A second convolutional layer (now using a step of 1 pixel) followed by another max-pooling operation result in 256 planes of $12\times$ 12 values. Three more convolutional layers were used and the resulting 256 planes of $12\times$ 12 values were flattened into a vector of $36,864$ values. Two fully connected layers are used to reduce these values to a single value, which was then passed through a sigmoid function to output an estimated probability.

Finally, the organ was classified as positive (i.e. displaying signs of the condition in question) or negative based the heat map(s) of patch probabilities. We experimented with two methods. The first method simply classifies patches by thresholding the patch probability at 0.5 and then classifies the organ by thresholding the proportion of patches classified as positive in the two images of that organ. The second method summarises heat-map statistics in the form of a 10-bin histogram of the patch probabilities and then classifies the organ using a support vector machine trained on such histograms.

The SVM outputs a signed score. The magnitude of this score can be taken as an indication of confidence in the classification. We can opt to reject any low confidence images, i.e. those with score magnitude below a threshold, and make classification decisions for the remaining images.

Pericarditis typically affects the appearance of most of an affected heart's surface. Therefore, we trained a heart patch classifier to classify heart patches into the following two classes: a positive class, consisting of all patches from hearts affected by periarditis, and a negative class, consisting of all patches from hearts unaffected by pericarditis. On the other hand, milk spots typically affect small localised regions of an affected liver's surface. Therefore, we trained a liver patch classifier to classify liver patches into the following two classes: a positive class consisting of patches centred near milk spot centres, and a negative class consisting of all other liver patches. Specifically, a liver patch was considered positive only if its central 48×48 -pixel portion contained the centre of a milk spot.

TABLE II: Reliability statistics for liver and heart classification

Condition	Raters	Over. hom.	Cat.	Marg. hom.	Sp. agr.
Pericarditis (n=501)	KY_1 v KY_2	$\chi^2 = 2.073$ p=0.3547	Neg	$\chi^2 = 0.000, p = 1.0000$	0.973
			Pos	$\chi^2 = 0.800, p = 0.3711$	0.949
			Inv	exact $p = 0.2891$	0.200
Pericarditis (n=230)	KY_1 v JW	$\chi^2 = 68.870$ p=0.0000 *	Neg	$\chi^2 = 45.082, p = 0.0000$ *	0.824
			Pos	$\chi^2 = 53.000, p = 0.0000$ *	0.694
			Inv	exact $p = 0.0703$	0.000
Milk spots (n=392)	KY_1 v KY_2	$\chi^2 = 54.069$ p=0.0000 *	Neg	$\chi^2 = 32.961, p = 0.0000$ *	0.882
			Pos	$\chi^2 = 46.621, p = 0.0000$ *	0.830
			Inv	$\chi^2 = 11.000, p = 0.0009$ *	0.000
Milk spots (n=164)	KY_1 v JW	$\chi^2 = 25.569$ p=0.0000 *	Neg	$\chi^2 = 15.207, p = 0.0001$ *	0.760
			Pos	$\chi^2 = 20.571, p = 0.0000$ *	0.863
			Inv	exact $p = 0.2500$	0.000

(a) (b) (c)

Fig. 4: Example annotations of (a) a heart, (b) a liver, and (c) milk spots.

A. Image Annotation

Instances of hearts and livers, both with and without pathology, were segmented by delineating their boundaries. These boundaries were then filled to create organ masks. In some cases the heart was only visible in one of the two images due to occlusion by lungs. Fig. 4(a) shows an example pluck in which the heart is annotated in each of the two views. Similarly, Fig. 4(b) shows a liver annotated in its two views. KY and TA provided the manual segmentations; TA was trained to delineate organs by JW and IK (a veterinary surgeon with more than 30 years experience in pig health, including detection of diseases and production). In order to efficiently obtain a large set of image pairs with accurately segmented organs we adopted a semi-automatic procedure making use of our previously developed method for automatic, simultaneous segmentation of the heart, lungs, diaphragm and liver, as well as the upper portion of the pluck [16]. The automatic method was judged by the human annotator to have produced an accurate segmentation of both the heart and liver for 183 plucks. The remaining organ instances were segmented manually.

Individual milk spots were annotated to enable development and validation of a milk spot detection algorithm. KY annotated the locations of all milk spots in a set of images that had been identified as containing liver affected by that condition. A tablet computer with stylus was used to mark the perceived centre of each milk spot. Fig. 4 shows detail from a liver region in which the centres of two milk spots have been marked (green dots).

B. Training and Validation

Data augmentation was used to increase the size and representativeness of the set of positive liver spot patches. First, we extracted a patch exactly centred on each manually marked milk spot and obtained three additional versions of that patch by rotating it 180 degrees and by flipping it vertically and horizontally. These operations preserved the vertical orientation of the patch, which is potentially important, given that the pluck hangs from a hook and is easily deformable. In addition, from each of these four versions of the patch we created nine additional versions by applying a translation randomly sampled from a two-dimensional normal distribution.

Organ classifications provided by KY_1 were used when compiling datasets for CNNs. We used a dataset of 450 livers of which 221 livers did not have any milk spots. The remaining 229 livers had a total of 1,748 milk spots that were manually marked, resulting in 6,753 positive patches. (Each milk spot centre results in multiple positive patches, given that patches are sampled at every 24 pixels horizontally and vertically). Data augmentation resulted in an additional 69,360 patches (i.e. an additional 40 patches per annotated milk spot, excluding some off-boundary cases), for a total of 76,113 positive patches. Furthermore, 2,324,160 negative liver patches were available. No data augmentation was used for heart patches because large numbers of positive patches can be readily sampled from heart images labelled as exhibiting signs of pericarditis. The dataset of 382 hearts contained 166 hearts without any signs of pericarditis and 216 hearts affected by pericarditis. This dataset yielded 215,498 positive patches and 168,209 negative patches.

Performance was evaluated using ten-fold stratified cross-validation. For each fold, three sets of patches were extracted from the pool of non-test organs: a balanced training set of 48,000 patches, a balanced validation set of 16,000 patches, and a uniformly sampled refinement set of 48,000 patches. It should be emphasised that these training, validation and refinement sets were all sampled from non-test data, so that at each fold 10% of the data were reserved for testing. At regular intervals during training (every 3 epochs) the partially trained model was used to classify the patches in the refinement set; misclassified refinement set patches were then moved to the training set, replacing randomly selected training patches. In this way the training set was regularly refreshed with patches that are difficult to classify. After each refinement step, discarded training patches and non-selected refinement patches were merged with the pool of unselected

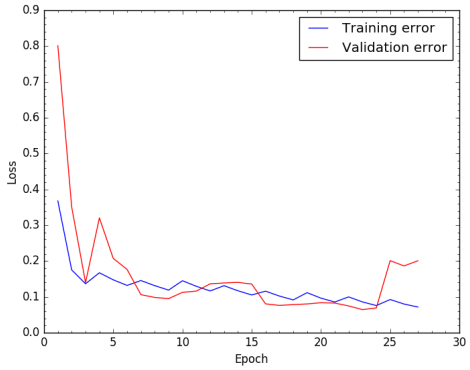


Fig. 5: Example training and validation loss curves using the training set refinement method.

non-test patches, and a refinement set was randomly drawn again. This refinement procedure helps to avoid overfitting and provides a mechanism for dealing efficiently with highly unbalanced data; rather than undersampling the dominant class at random, it samples disproportionately those examples that are misclassified. **Similar bootstrap methods for mining 'hard' samples are often used in the computer vision literature to iteratively refine the training set either by replacing part of it or augmenting it, e.g., [27], [28], [29], [30]. Fig. 5 shows training and validation loss for an example liver fold using this method.**

CNN patch classifiers were trained for 27 epochs (with a refinement step every 3 epochs) and the model stored after each epoch. The model that achieved the lowest validation error was then selected. Training was done by mini-batch gradient descent, using mini-batches of 48 patches and cross-entropy loss. The learning rate at each epoch was determined by $\eta = 0.01/(1 + 0.2\times$ epoch $)$ and a weight decay of 0.0005 was used for regularisation. SVM organ classifiers with radial basis function kernels were trained using the classifications provided by KY_1. The SVM penalty parameter C and kernel coefficient γ were tuned via nested 5-fold cross-validation.

VI. Results

Table III reports confusion matrices obtained when comparing automatic classification using SVM with that of a human expert (KY_1). Table IV provides statistical analysis of these confusion matrices. The specific agreements for each class are given along with the result of McNemar tests for marginal homogeneity. In the case of heart classification, the agreement was better than in the inter-observer study. Furthermore, the test for lack of marginal homogeneity indicated that the system was using the classes with similar propensity to the human expert. In the case of liver classification, the agreement was comparable with the inter-observer study but the classes were assigned with differing propensity. The relatively few instances classified as *invalid* were excluded from this analysis for simplicity.

Fig. 6 visualises the SVM score distributions as box and whisker plots. The box extends from the lower to upper quartile values with a line at the median. The class scores

TABLE III: Confusion matrices of milk spot and pericarditis detection.

(a) Milk spots

		Predicted	
		Neg	Pos
True	Neg	191	30
	Pos	52	177

(b) Pericarditis

		Predicted	
		Neg	Pos
True	Neg	155	11
	Pos	16	200

TABLE IV: Reliability statistics when comparing automatic classification with an expert (KY_1). SA denotes specific agreement.

Condition	McNemar test	Class	SA
Pericarditis (n=382)	$\chi^2 = 0.926, p = 0.3359$	Neg	0.920
		Pos	0.937
Milk spots (n=450)	$\chi^2 = 5.902, p = 0.0151$ *	Neg	0.823
		Pos	0.812

are well separated, especially for heart classification. The default classification rule was to decide the class based on thresholding the score at zero (i.e. the sign of the score); false positive and false negative errors can instead be traded off against each other by varying the decision threshold. Fig. 7 shows ROC curves thus obtained. The area under the curve (AUC) was 0.958 for heart classification and 0.893 for liver classification. Also shown are the curves obtained when, instead of an SVM, classification was performed by thresholding the proportion of positive patches contained in the heat map. In the case of heart classification, this gave a lower AUC than the SVM of 0.923, and the SVM curve dominates except at low false postive rates where the two methods are comparable. In the case of liver classification it gave an AUC of 0.892 and the ROC curves cross over.

We used heat maps to visualise patch probabilities as grey levels ranging from black (denoting zero probability) to white (denoting a probability of one). Fig. 9 shows some example images along with their manual liver and liver spot annotations (ground truth), and their automatically generated liver spot heat maps. Two livers that were classified correctly (as negative and positive respectively) and two that were misclassified are shown. Fig. 10 similarly shows examples of heat maps for four hearts. The two misclassified hearts shown are instances that were misclassified with high confidence (i.e., with high magnitude SVM scores). Fig. 8 shows how, as the score magnitude threshold is increased, misclassified organs are disproportionately rejected.

Finally, we report results at the patch level for liver. As explained previously, when the marked centre of a milk spot falls within the central portion ($48\times$ 48 pixels) of a given patch, that patch is considered positive. Thresholding the patch probabilities acts as a liver spot detector. Table V shows the confusion matrix for classification of the 2,330,913 patches located within the 450 livers (when probability was thresholded at 0.5). The true positive rate (recall) was 0.843 and the false positive rate was 0.016. Precision was 0.132.

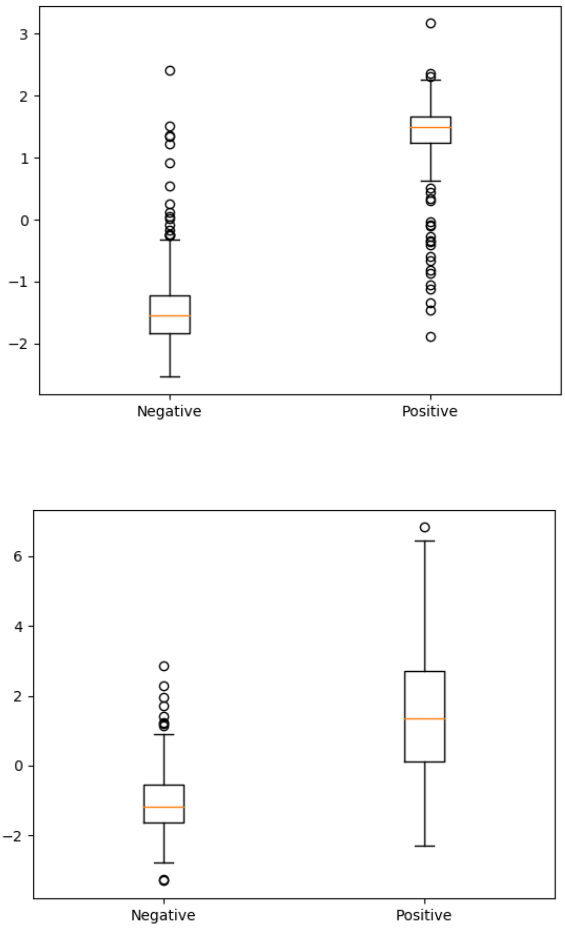
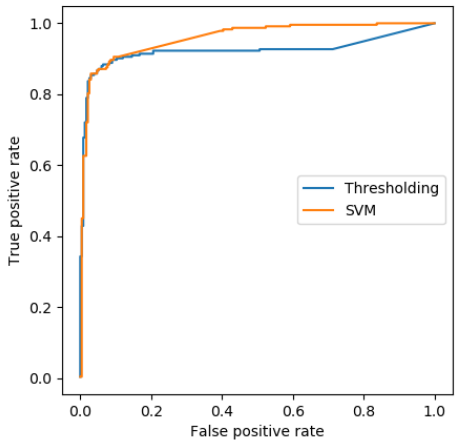


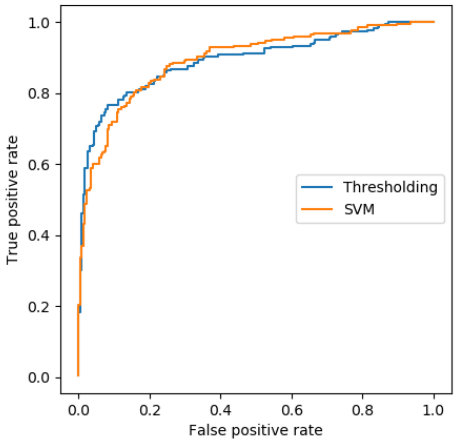
Fig. 6: Box and whiskers plots of SVM scores (distance to SVM decision boundary) for heart classification (top) and liver classification (bottom).

TABLE V: Confusion matrix for liver patch classification.

		Predicted	
		Neg	Pos
True	Neg	2286801	37359
	Pos	1062	5691



(a)



(b)

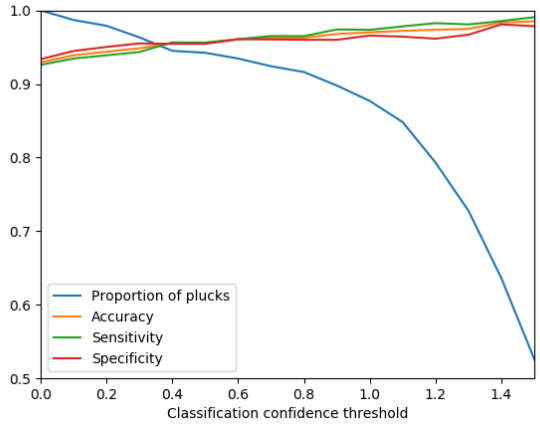
Fig. 7: ROC curves. (a) Heart (pericarditis) classification. (b) Liver (milk spots) classification

VII. Analysis and Discussion

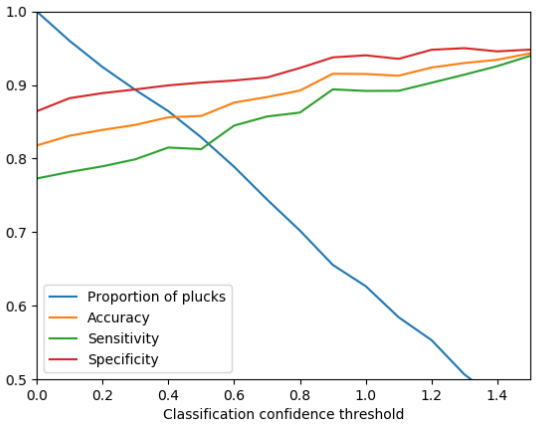
A. Analysis of Classification Results

Milk spots appear as blobs of milky colour that vary in appearance considerably. Liver surface regions free of milk spots also vary, exhibiting textures and specular reflections that may mimic milk spots in appearance (see Fig. 12). Fig. 9 shows an example false positive classified with high confidence, a type of misclassification that occurs mostly when there is surface texture that the CNN mistakes for milk spots; in this example the liver was covered with froth that accumulated in bright blobs (see detail in Fig. 11(a)). Most false negatives were livers with very few milk spots, as in the example in Fig. 9 which has a single annotated milk spot. At the patch level, liver spot patches were detected with high recall but relatively low precision (Table V). A few problematic livers gave rise to large numbers of false positive patches resulting in the low overall patch precision. Nevertheless, patch probabilities (heat maps) were sufficient to enable good organ classification (AUC $= 0.89$) comparable with the human expert. Specific agreements were 0.82 (negative class) and 0.81 (positive class) whereas inter-rater specific agreements between the human experts were 0.76 (negative class) and 0.86 (positive class).

Variability in appearance of hearts with and without pericarditis is illustrated in Fig. 13; hearts free from pericarditis generally have a smoother, shinier surface. Fig. 11 shows detail from images of the misclassified hearts in Fig. 10. The false negative shown here had an unusually shiny surface for a heart affected by pericarditis. The false positive displayed a relatively milky surface for a heart that is healthy. It also had an incision with associated clotted blood accidentally inflicted during handling of the pluck which presumably contributed to this misclassification. Overall, heart classification had high accuracy (AUC $= 0.96$). Specific agreements were 0.92 (negative class) and 0.94 (positive class) which compare very



(a)



(b)

Fig. 8: Plots of classification measures and proportion of classified plucks against the SVM score magnitude threshold used to reject low confidence images. (a) Heart. (b) Liver.

favorably with the inter-rater specific agreements between the human experts of 0.82 (negative class) and 0.69 (positive class). Sometimes hearts remain within their pericardial sac (the healthy normal membrane around the heart) as abattoir staff may not incise it. Such cases can be completely normal but for the purposes of this study both observers classed them as pericarditis although they obviously do not look like the typical pericarditis cases.

In summary, automatic heart classification was in closer agreement with the expert on which it was trained than the two experts were with each other. Automatic liver classification performed at a level similar to inter-expert agreement. In both cases, classification agreement was somewhat lower than *intra*-expert agreement. However, these intra-expert agreements constitute a very demanding human reference. The expert pathologists were not time-restricted when carefully identifying the pathologies of interest for this research. In a real abattoir situation, a human inspector would have significantly less time to visually inspect an offal, which would be expected to lead to significantly lower inter-rater and intra-rater agreements.

We compared the use of SVM heat map classifiers to simply thresholding the proportion of positive patches. For hearts, SVM gave a higher AUC whereas for livers it gave no gain in performance (Fig. 7). A measure of classification confidence could be used to decide which organs to classify automatically and which organs to refer for manual assessment by a human inspector. Fig. 8(b) shows that, using SVM score magnitude as confidence measure, rejection of cases with confidence below 0.3 resulted in 10.7% of livers being rejected (for manual assessment). The remaining 89.3% of livers were classified with an overall accuracy of 84.6%. This is the same level of accuracy obtained in the intra-expert study albeit without a rejection option (see Table I(a)).

B. Detecting different pathologies

We developed a methodology that enabled automated classification of two pathologies of interest in pig carcasses. The two pathologies, namely liver milk spots and pericarditis, as well as being significant from a public health perspective (e.g. milk spots are the most common pathology identified on pig carcasses [3]) were also used as a case in point of spot-like and diffuse pathologies, respectively. In fact, many other pathologies of interest on pig offal, for example inflammation of the pleura, can be considered spatially diffuse, whereas several types of lung inflammation, such as those resulting from viral infection, may lead to multi-focal pathology[2] [5]. Therefore, the framework can easily be extended to other pathologies based on provision of labelled image sets for training, validation and testing.

The main difficulty associated with training machine learning models to detect affected regions on the surface of the organs is the fact that both negative (i.e. normal) and positive regions vary widely in appearance. As a result, a large amount of data is needed to successfully train the models to distinguish positive and negative regions, meaning that many images of organs have to be manually labelled by veterinary experts as to the presence of the pathologies of interest. This process is time-consuming and expensive, and could potentially be compounded by the need to not only label organs but also manually segment the lesions on their surfaces, to provide positive training examples. In the case of pericarditis we chose to dispense with manual segmentation of lesions, instead opting to use the entire surface of affected hearts as source of positive training data. This approximation enabled high classification performance while avoiding the costs associated with manual segmentation of lesions. In the case of liver milk spots, given their extremely localised nature, we chose to ask experts to mark the centres of milk spots, an approach which was much less time-consuming than fully segmenting milk spots by hand. The width of a small number of milk spots was measured to help determine the appropriate size for the individual image patches that were analysed, as well as the overlap between adjacent patches. The application of machine learning to classify pathologies that are more spatially localised than pericarditis but not as localised as milk spots

[2]https://www.pig333.com/pathology-atlas/

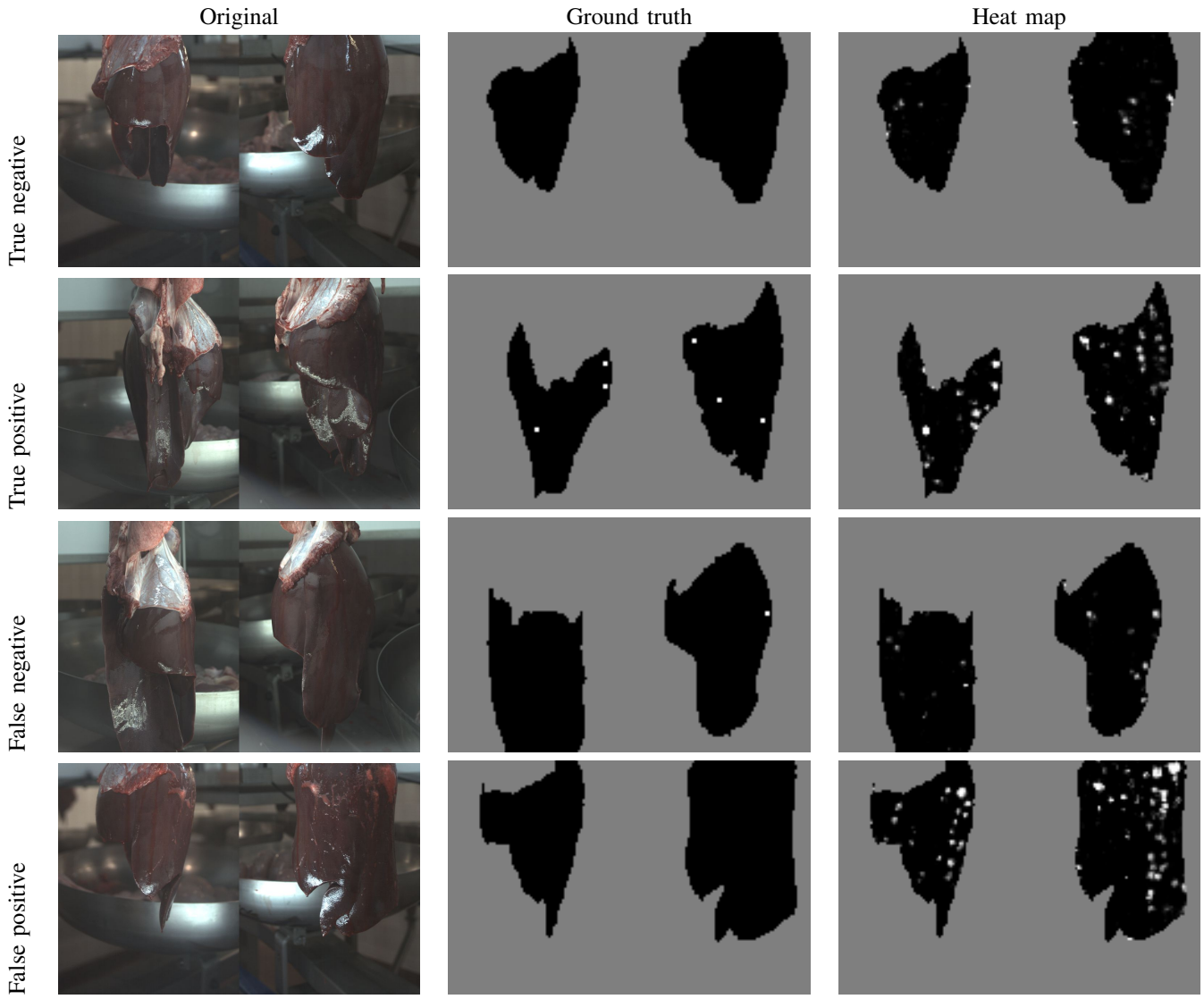


Fig. 9: Example heat maps obtained through liver patch classification with corresponding ground truth and original images.

might benefit from annotation of lesions' spatial extent in the form of closed polygons, for example.

C. Image capture

Successful analysis relies on images that capture the surface of organs with fine texture detail. This in principle requires not only high resolution cameras but also exposure with sufficient depth of field, to keep as much of the surface of the organs within focus. We found this to require a fine balance, as large depths of field are typically attained by a combination of small aperture and sufficient distance from the camera to the object (possibly compensated by zooming). In practice, our image capture equipment had to be installed in close proximity to the human inspectors, precluding the use of strobe lighting and therefore implying the use of relatively large apertures. In addition, the requirement to keep lighting and cameras inside of a canopy structure caused the cameras to be relatively close to the offal. Another technical obstacle we faced was the presence of large specular reflections from the lighting sources on the surface of the organs, which appear in the images as large areas of saturated luminance, containing no textural information. We minimised this by fitting both light sources and cameras with polarising filters.

D. Limitations

This paper explored automatic classification of segmented organs. Currently, automatic organ segmentation operates well under fixed lighting conditions but is not yet robust enough for deployment; future work in that direction could usefully explore structure learning [31] and end-to-end learning of segmentation maps [32]. It is likely that using more cameras to obtain more complete coverage of the offal could improve accuracy of automated identification albeit at the expense of additional computation and installation costs. Future work could investigate global organ descriptors other than the patch histograms used in this paper, perhaps incorporating morphological information. Some misclassified instances had unusual appearance suggesting that even larger, more repre-

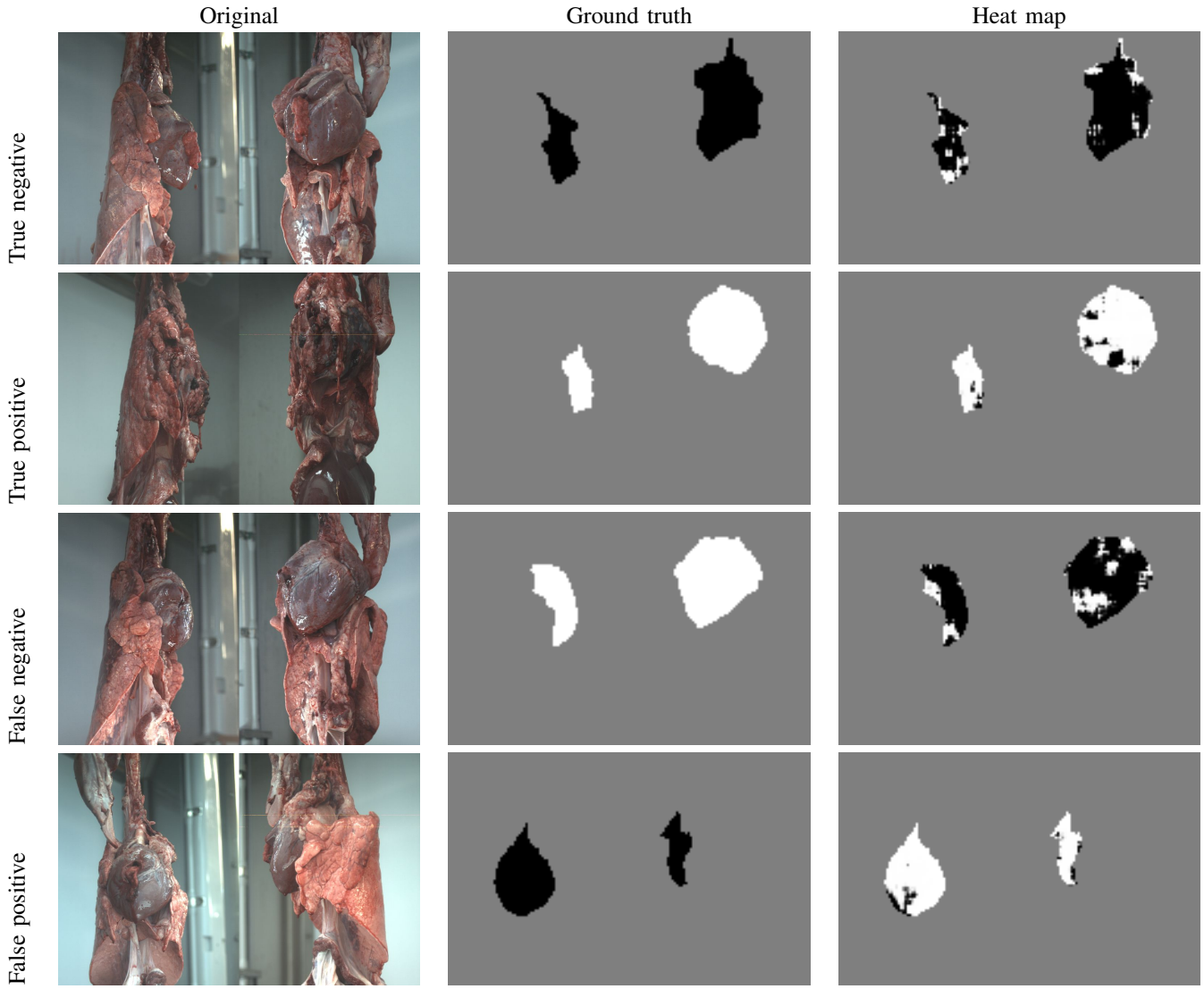
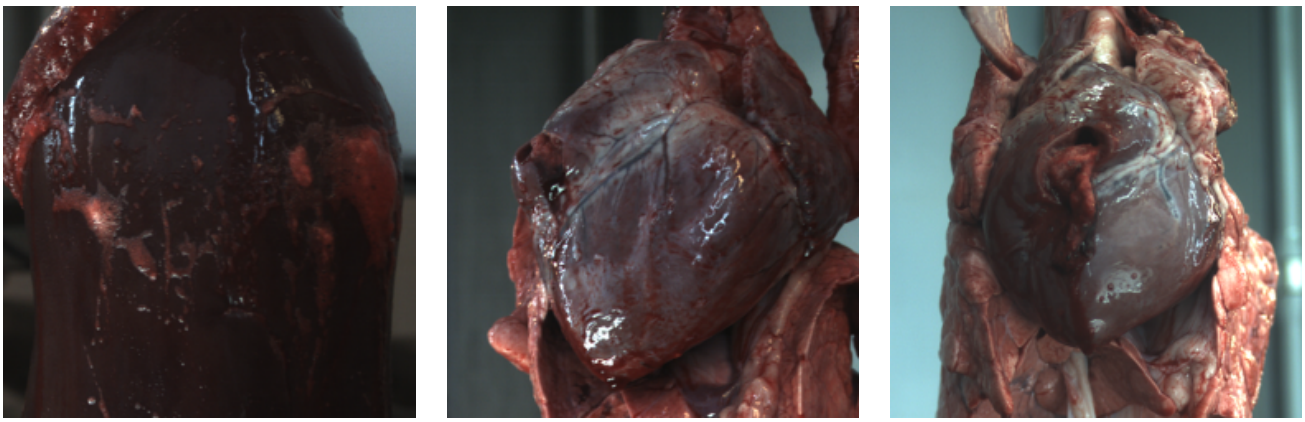
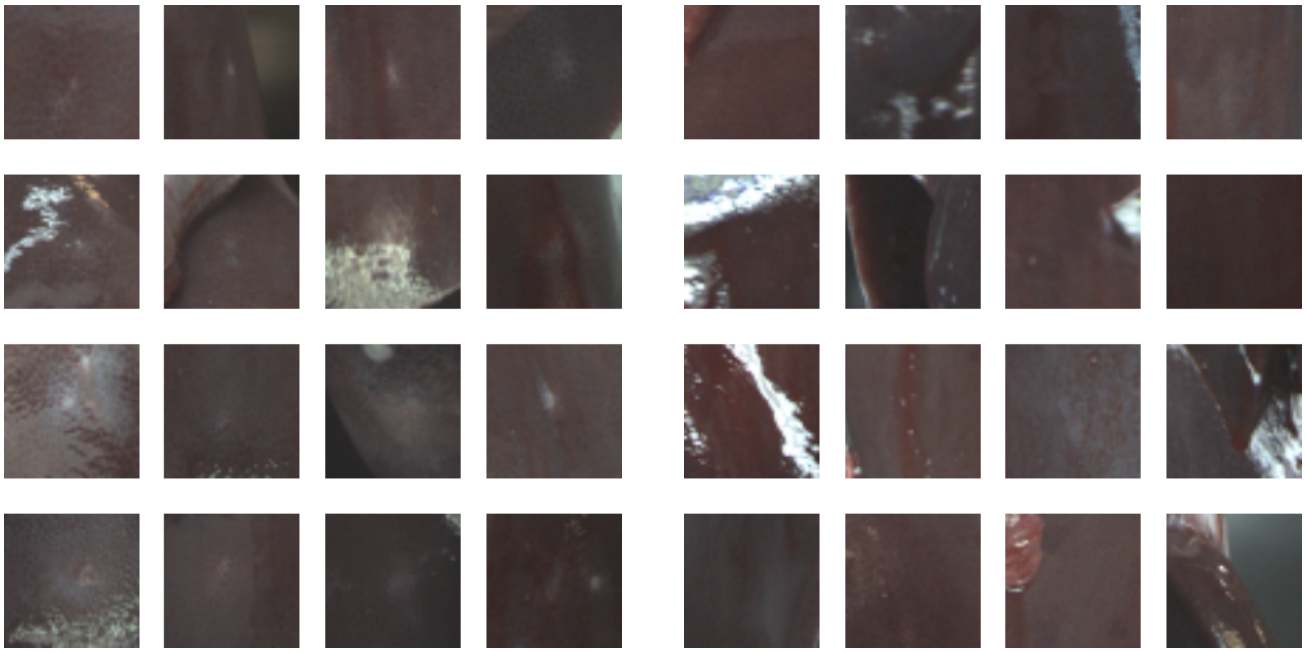


Fig. 10: Example heat maps obtained through heart patch classification with corresponding ground truth and original images.



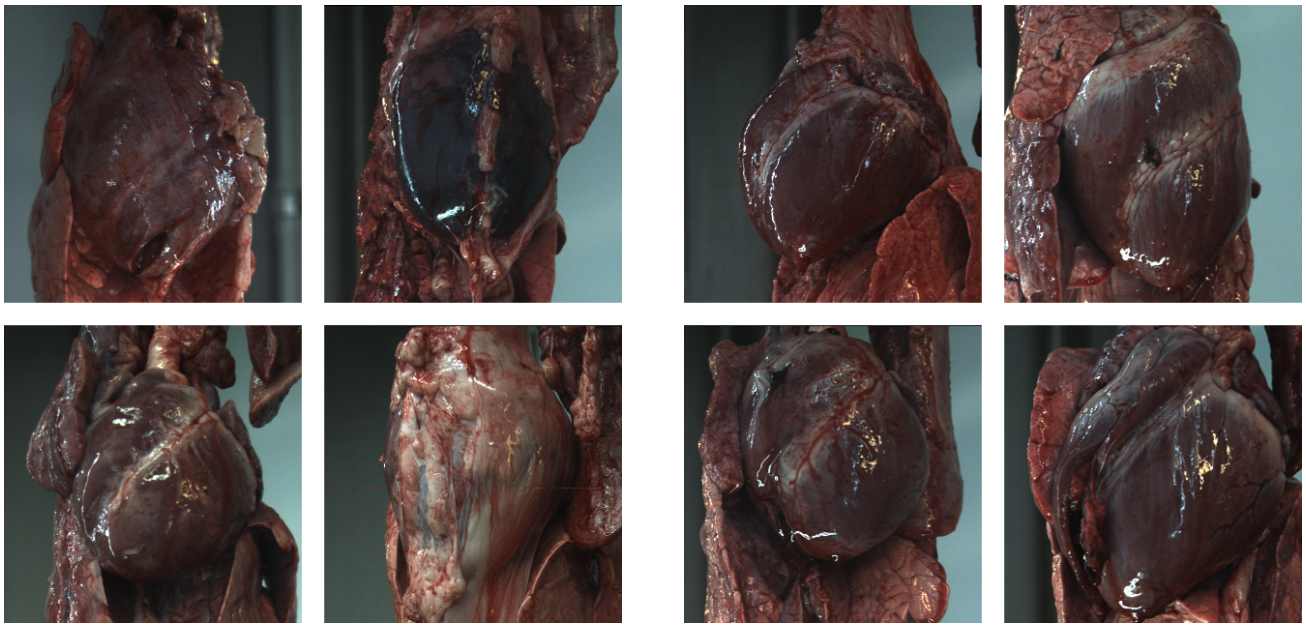
(a) (b) (c)

Fig. 11: (a) Detail of false positive liver shown in Fig. 9. Detail of (b) false negative and (c) false positive hearts shown in Fig. 10.



(a) (b)

Fig. 12: Variability of appearance of (a) liver milk spots and (b) similarly sized regions without milk spots. (Images have been gamma-adjusted for visualization)



(a) (b)

Fig. 13: Examples of hearts (a) with pericarditis and (b) without pericarditis.

sentative labelled training datasets would be likely to improve performance further. **As the size of training sets grows, bootstrapping methods that select more efficiently which locations to consider could be worth adopting [33].**

VIII. Conclusion

There has been a recent trend to move to visual-only inspection of pig carcasses, to minimize the risk of cross contamination between carcasses that may arise from palpation or incision [1]. A system using methods such as those presented here has the potential to (i) comply with this trend, (ii) overcome limitations that arise from human subjectivity, lack of clarity in evaluating the pathologies and significant inter-rater disagreement, and (iii) ultimately provide a new gold standard for pathology. Although a first necessary step in the development of a practical, automated system for screening pig pathologies at abattoir, it holds promise, even if it is to be used as the initial screening for affected offal that could then be examined in greater detail by a meat inspector.

IX. Acknowledgments

This research was undertaken as part of the BBSRC / Innovate UK Project *Automated screening for pathologies at abattoir through computer vision based inspection of pig carcasses* in collaboration with Tulip Ltd. and Hellenic Systems Ltd. The authors were funded by BBSRC grants BB/L017385/1 (Amaral and Kyriazakis, Newcastle University) and BB/L017423/1 (McKenna, University of Dundee). The project also received funding from Tulip Ltd., Hellenic Systems Ltd., and Innovate UK.

We are grateful to Katharine Yuill and Jake Waddilove for manually annotating the images used in this study, and to Thomas Plötz for his input to this project. We are also grateful to colleagues at Tulip Ltd and Hellenic Systems Ltd for their continual support throughout this research. Importantly, Tulip Ltd provided access to their abattoir facility, and Hellenic Systems Ltd designed and developed the image capturing system installed within the abattoir.

References

[1] EFSA Panels on Biological Hazards (BIOHAZ), on Contaminants in the Food Chain (CONTAM), and on Animal Health and Welfare (AHAW), "Scientific opinion on the public health hazards to be covered by inspection of meat (swine)," *EFSA Journal*, vol. 9, no. 10, p. 2351 (198 pp.), 2011, doi:10.2903/j.efsa.2011.2351.

[2] S. S. Nielsen, G. B. Nielsen, M. J. Denwood, J. Haugegaard, and H. Houe, "Comparison of recording of pericarditis and lung disorders at routine meat inspection with findings at systematic health monitoring in Danish finisher pigs," *Acta Veterinaria Scandinavica*, vol. 57, no. 1, p. 18, Mar 2015. [Online]. Available: https://doi.org/10.1186/s13028-015-0109-z

[3] H. R. Holt, P. Alarcon, M. Velasova, D. U. Pfeiffer, and B. Wieland, "BPEX pig health scheme: a useful monitoring system for respiratory disease control in pig farms?" *BMC Veterinary Research*, vol. 7, no. 1, p. 82, Dec 2011. [Online]. Available: https://doi.org/10.1186/1746-6148-7-82

[4] T. Steinmann, T. Blaha, and D. Meemken, "A simplified evaluation system of surface-related lung lesions of pigs for official meatinspection under industrial slaughter conditions in Germany," *BMC Veterinary Research*, vol. 10, no. 1, p. 98, Apr 2014.

[5] J. J. Zimmerman, L. A. Karriker, A. Ramirez, K. J. Schwartz, and G. W. Stevenson, Eds., *Diseases of Swine*. Hoboken NJ: Wiley-Blackwell, 2012.

[6] J. Buttenschon, N. F. Friis, B. Aalbaek, J. T. K., T. Iburg, and J. Mousing, "Microbiology and pathology of fibrinous pericarditis in Danish slaughter pigs." *Journal of Veterinary Medicine Series A*, vol. 44, no. 5, pp. 271–80, 1997.

[7] Y. Le Cun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[8] J. Ma, D.-W. Sun, J.-H. Qu, D. Liu, H. Pu, W.-H. Gao, and X.-A. Zeng, "Applications of computer vision for assessing quality of agri-food products: A review of recent research advances," *Critical Reviews in Food Science and Nutrition*, vol. 56, no. 1, pp. 113–127, 2016.

[9] C. Craigie, E. Navajas, R. Purchas, C. Maltin, L. Bnger, S. Hoskin, D. Ross, S. Morris, and R. Roehe, "A review of the development and use of video image analysis (VIA) for beef carcass evaluation as an alternative to the current EUROP system and other subjective systems," *Meat Science*, vol. 92, no. 4, pp. 307 – 318, 2012.

[10] Z. Xiong, D.-W. Sun, H. Pu, W. Gao, and Q. Dai, "Applications of emerging imaging techniques for meat quality and safety detection and evaluation: A review," *Critical Reviews in Food Science and Nutrition*, vol. 57, no. 4, pp. 755–768, 2017.

[11] R. Vanderhasselt, M. Sprenger, L. Duchateau, and F. Tuyttens, "Automated assessment of footpad dermatitis in broiler chickens at the slaughter-line: evaluation and correspondence with human expert scores," *Poultry Science*, vol. 92, no. 1, pp. 12–18, 2013.

[12] A. Jørgensen, T. Moeslund, and E. Mølvig Jensen, "Detecting gallbladders in chicken livers using spectral analysis," in *Machine Vision for Animals and their Behaviour (BMVC workshop)*, 2015, pp. 2.1–2.8.

[13] A. Jørgensen, J. Fagertun, and T. Moeslund, "Diagnosis of broiler livers by classifying image patches," in *Scandinavian Conference on Image Analysis (SCIA)*. Springer, 2017, pp. 374–385.

[14] Y. Chen and S. C. Wang, "Poultry carcass visceral contour recognition method using image processing," *The Journal of Applied Poultry Research*, vol. 27, no. 3, pp. 316–324, 2018.

[15] M. Philip Philipsen, J. Velling Dueholm, A. Jørgensen, S. Escalera, and T. Moeslund, "Organ segmentation in poultry viscera using RGB-D," *Sensors*, vol. 18, p. 117, 01 2018.

[16] S. McKenna, T. Amaral, T. Plötz, and I. Kyriazakis, "Multi-part segmentation for porcine offal inspection with auto-context and adaptive atlases," *Pattern Recognition Letters*, vol. 112, pp. 290 – 296, 2018.

[17] K. Chao, Y.-R. Chen, H. Early, and B. Park, "Color image classification systems for poultry viscera inspection," *Applied Engineering in Agriculture*, vol. 15, no. 4, p. 363, 1999.

[18] J. Ibarra, Y. Tao, L. Newberry, and Y. Chen, "Learning vector quantization for color classification of diseased air sacs in chicken carcasses," *Transactions of the ASAE*, vol. 45, no. 5, p. 1629, 2002.

[19] Y. Tao, J. Shao, K. Skeeles, and Y. R. Chen, "Detection of splenomegaly in poultry carcasses by UV and color imaging," *Transactions of the ASAE*, vol. 43, no. 2, p. 469, 2000.

[20] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, p. 29, 2016.

[21] G. Litjens, C. I. Sanchez, N. Timofeeva, and J. A. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports*, vol. 6, May 2016.

[22] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A comprehensive computer-aided polyp detection system for colonoscopy videos," in *International Conference on Information Processing in Medical Imaging*, 2015, pp. 327–338.

[23] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. Lau, and C. C. Poon, "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 41–47, 2017.

[24] L. K. Ferris, J. A. Harkes, B. Gilbert, D. G. Winger, K. Golubets, O. Akilov, and M. Satyanarayanan, "Computer-aided classification of melanocytic lesions using dermoscopic images," *Journal of the American Academy of Dermatology*, vol. 73, no. 5, pp. 769–776, 2015.

[25] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[27] S. J. McKenna, "Automated analysis of papanicolaou smears," Ph.D. dissertation, University of Dundee, 1994.

[28] K. K. Sung, "Learning and example selection for object and pattern detection," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1996.

[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[30] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2010, pp. 1030–1037.

[31] S. Manivannan, J. Li, W. and. Zhang, E. Trucco, and S. J. McKenna, "Structure prediction for gland segmentation with hand-crafted and deep convolutional features," *IEEE Transactions on Medical Imaging*, 2017, doi: 10.1109/TMI.2017.2750210.

[32] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, Apr. 2017.

[33] O. Canévet and F. Fleuret, "Efficient sample mining for object detection," *JMLR: Workshop and Conference Proceedings*, vol. 39, pp. 48–63, 2014.